Управление разно-структурированными большими данными

Брюхов Дмитрий, Вовченко Алексей

dbriukhov@ipiran.ru

к.т.н., с.н.с.

ИПИ ФИЦ ИУ РАН

Общая информация

- Курс для магистров ВМК
- Рассматривается специальный вид стека для параллельных архитектур оперирования данными в аналитических приложениях Big Data
- Центральной в таком архитектурном стеке является парадигма программирования, называемая map-reduce.
- Такие архитектуры поддерживают разно-структурированные данные, которые могут быть представлены в разнообразных моделях данных (структурированных, слабоструктурированных, неструктурированных).
- В курсе рассматриваются основные идеи и подходы параллельных архитектур оперирования разно-структурированными данными.

Общая информация (2)

- Рассматриваются вопросы реализации различных алгоритмов в среде map-reduce (таких как матрично-векторное умножение, поддержка SQL-подобных операций и операций реляционной алгебры).
- Map-reduce программирование в курсе изучается применяя собственно API map-reduce Hadoop'а наряду с декларативными языками над Hadoop'ом (такими как PigLatin, Hive, Jaql).
- В комбинации с Hadoop'ом в курсе рассматриваются базы данных NoSQL (такие как HBase).
- Подходы к интеграции Hadoop'а в хранилище данных также рассматриваются.
- Рассмотрение методов применения аналитики данных над Hadoop'ом также является частью курса.

Содержание Курса

- Проблематику управления большими данными
- архитектура Hadoop, парадигма map-reduce
- Алгоритмы известных операций в map-reduce (матрично-векторное умножение, реляционные операции и др.)
- Разработка собственных приложений Map-Reduce
- Теория сложности алгоритмов для map-reduce
- Языки программирования высокого уровня над Hadoop: Pig, Hive, JAQL
- Базы данных нетрадиционных моделей данных Not Only SQL (NoSQL) и их связь с Hadoop (на примере HBase)
- Интеграция Hadoop с традиционными хранилищами данных (базами данных и warehouse)
- Методы анализ текстов над Hadoop

Введение в курс. Основные понятия курса. Введение в проблематику управления большими данными.

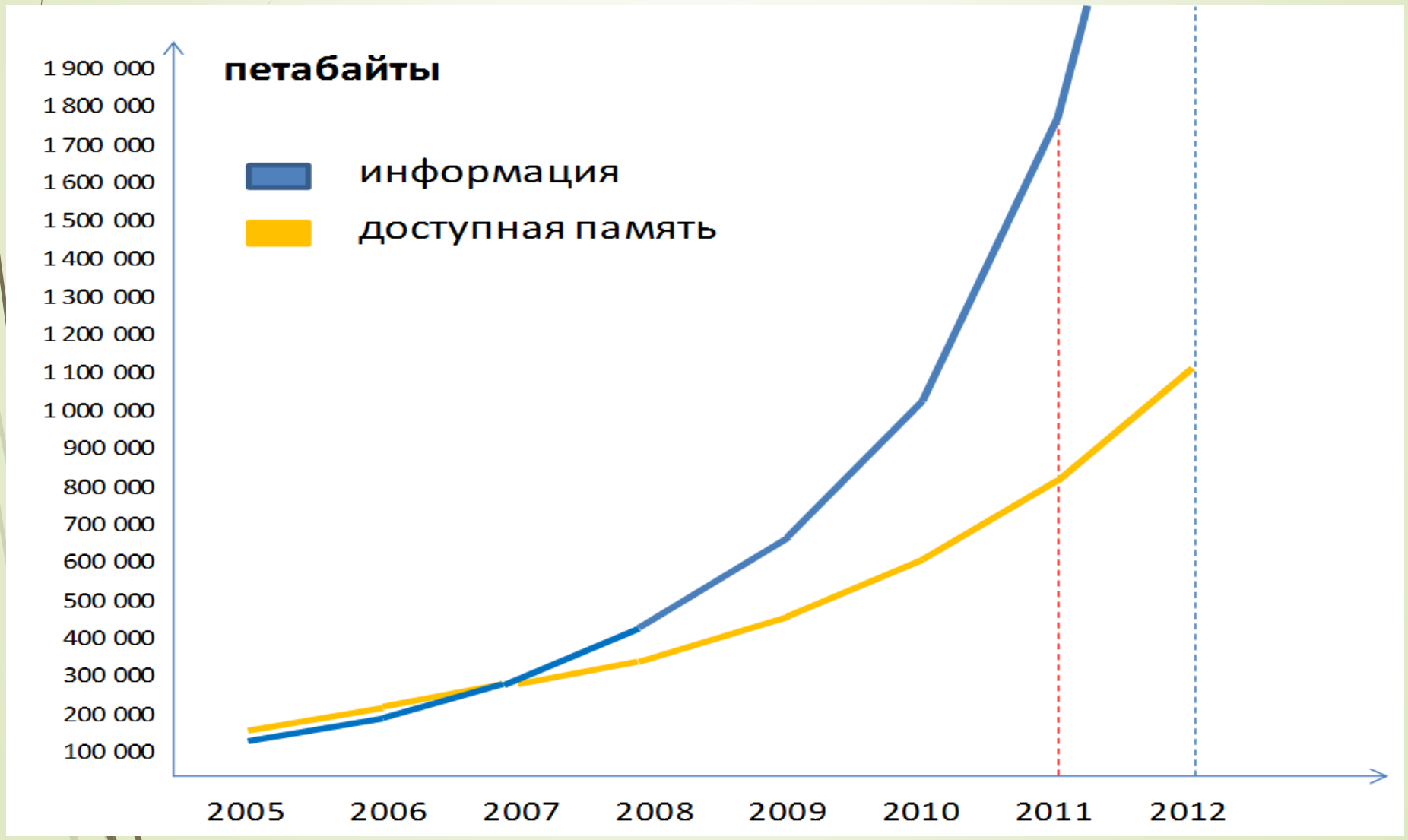
План

- Рост объема данных и типов данных
- Big Data как революционная технология
- Что такое большие данные?
- Кому нужны большие данные? (проекты и внедрения)
- Технологии больших данных
- Недостатки традиционных архитектур
- Платформа больших данных

Рост объема и разнообразия данных

- **Интернет:** web logs, facebook, twitter, maps, blogs, etc.
- **Финансы:** анализ данных большого объема для выявления тенденций и углубленного проникновения в состояние дел
- **Здравоохранение:** огромные данные о пациентах, лекарствах и способах лечения
- **Вселенная:** телескоп Хаббла обнаруживает многие сотни галактик, каждая с миллиардами звезд
- **Биоинформатика:** от данных о 3.3 млрд. оснований нуклеиновых кислот в геноме человека до огромного числа протеиновых последовательностей и анализа их поведения
- **Библиотеки:** всего 20 терабайт текстовых данных содержится в Библиотеке Конгресса США (130 млн. единиц хранения (книги, фото, карты и др.)

Рост цифровых данных и доступной памяти



Статистика объема данных

IDC:

ежегодно объемы хранимой информации вырастают на 40%

.

В 2012 г. во всех странах будет сгенерировано 2,43 Зеттабайт (1 ЗБ = около 1 млрд Гб),
это более чем в 2 раза превосходит объем информации в цифровом виде в 2010 г. (1,2 ЗБ).

Данные поступают из все большего количества источников
Данные не структурированы

Статистика объема данных (2)

- Каждый день в мире производится 2,5 квинтильона (10^{18}) байтов данных. 90% данных созданы за последние два года.
- Каждый час Wal-Mart совершает 1 миллион сделок, пополняя базу данных на 2,5 петабайта (10^{15})- в 170 раз больше объема данных Библиотеки Конгресса США.
- Объем отправлений, доставляемых американской Почтовой службой за один год, равен 5 петабайтам, а Google обрабатывает такой же объем данных всего за один час.
- Суммарный объем всей существующей на земле информации составляет несколько больше одного зеттабайта (10^{21}).
- Бизнесу во всем мире необходимо собрать, организовать, проанализировать все эти данные для достижения конкурентных преимуществ в конкретной сфере рынка, развития более эффективных деловых операций, увеличения удовлетворения запросов потребителей, выявления новых возможностей

Источники больших данных

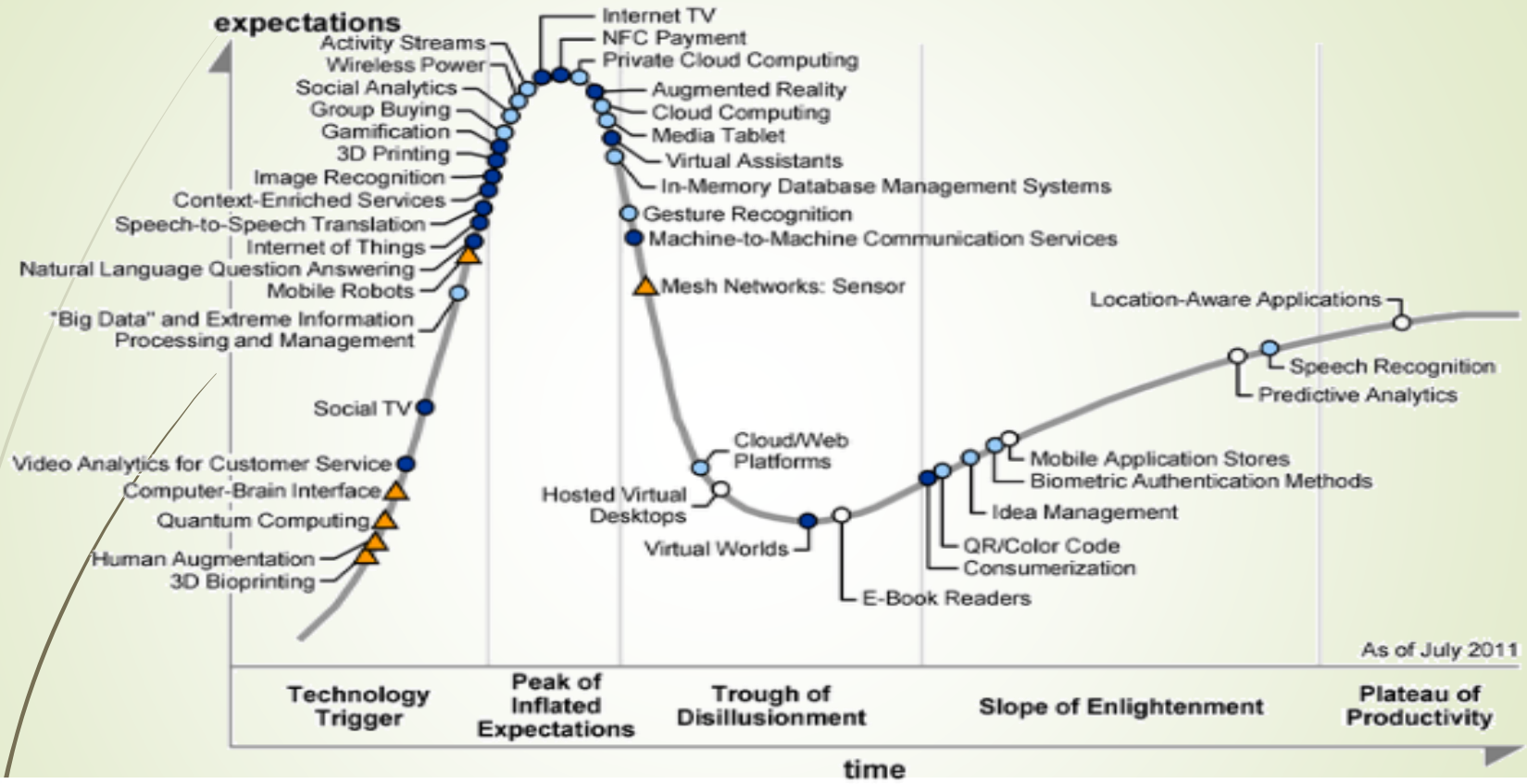
- Социальные сети и их данные
- Данные от измерительных устройств
- Журналы доступа пользователей веб-сайтов
- Сенсорные сети
- Тексты и документы из Интернета
- Научные данные (астрономия, геном человека, исследования атмосферы, биохимия, биология)
- Данные министерства обороны
- Медицинские наблюдения
- Фото- и видео-архивы
- Данные электронной коммерции
- …

Рекламный цикл Гартнера

Пять различных фаз, составляющих процесс развития любой новой технологии, включают:

- **Появление технологии.** Технологический прорыв, публичная демонстрация, выпуск продукта, или другое событие.
- **Пик завышенных ожиданий.** Фаза сверх-энтузиазма и нереалистичных прогнозов, подогреваемых лидерами технологии.
- **Впадина разочарования**. Период, в котором технология становится немодной и пресса оставляет эту тему.
- **Скат просветления**. Целенаправленные эксперименты и солидная, тяжелая работа все возрастающего числа различных организаций приводят к истинному пониманию применимости, рисков и преимуществ технологии.
- **Плато реализации**. Реальные достоинства технологии демонстрируются и воспринимаются в мире. Инструменты и методологии становятся все более стабильными.

Технология Big Data (2011)



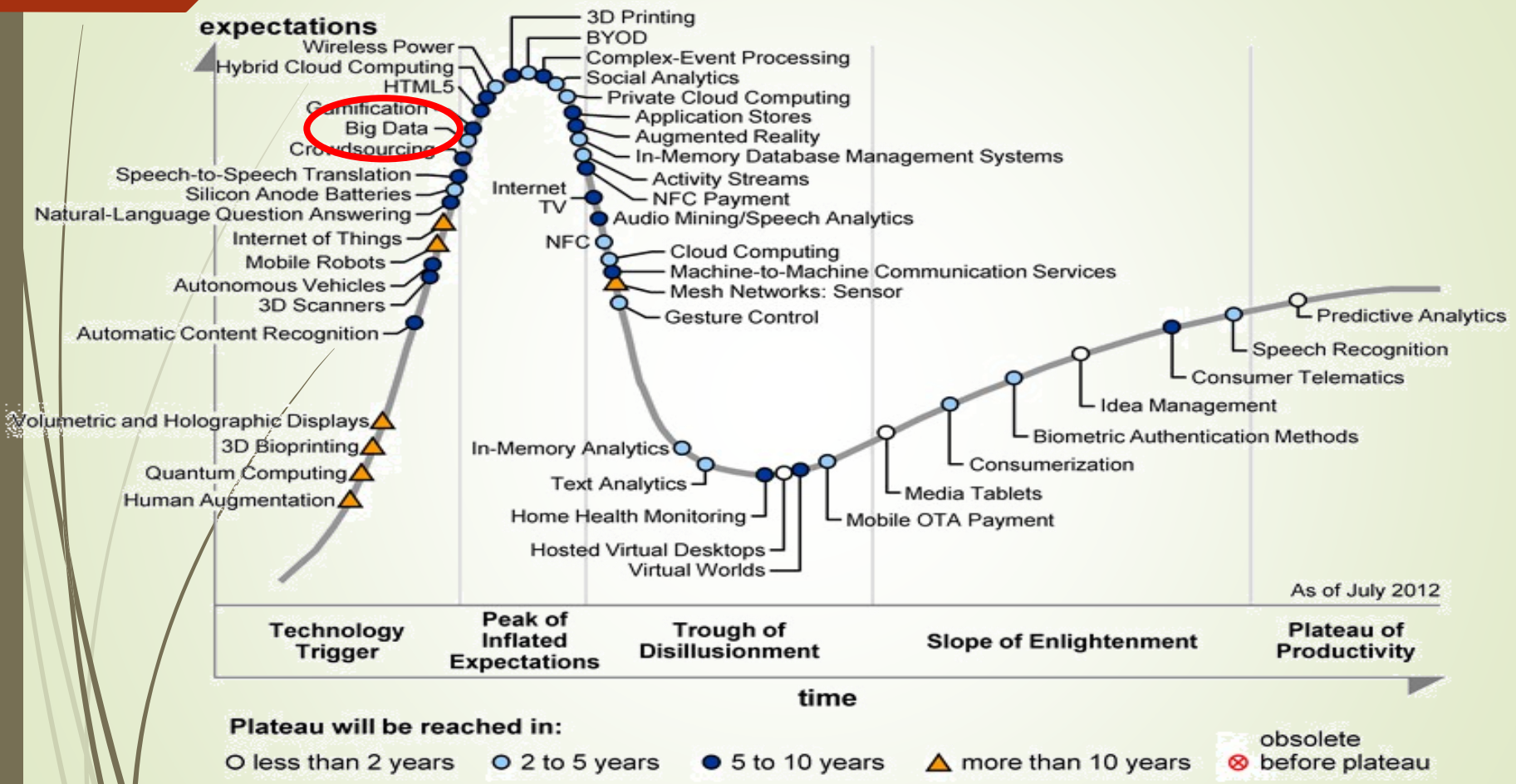
News

Years to mainstream adoption:

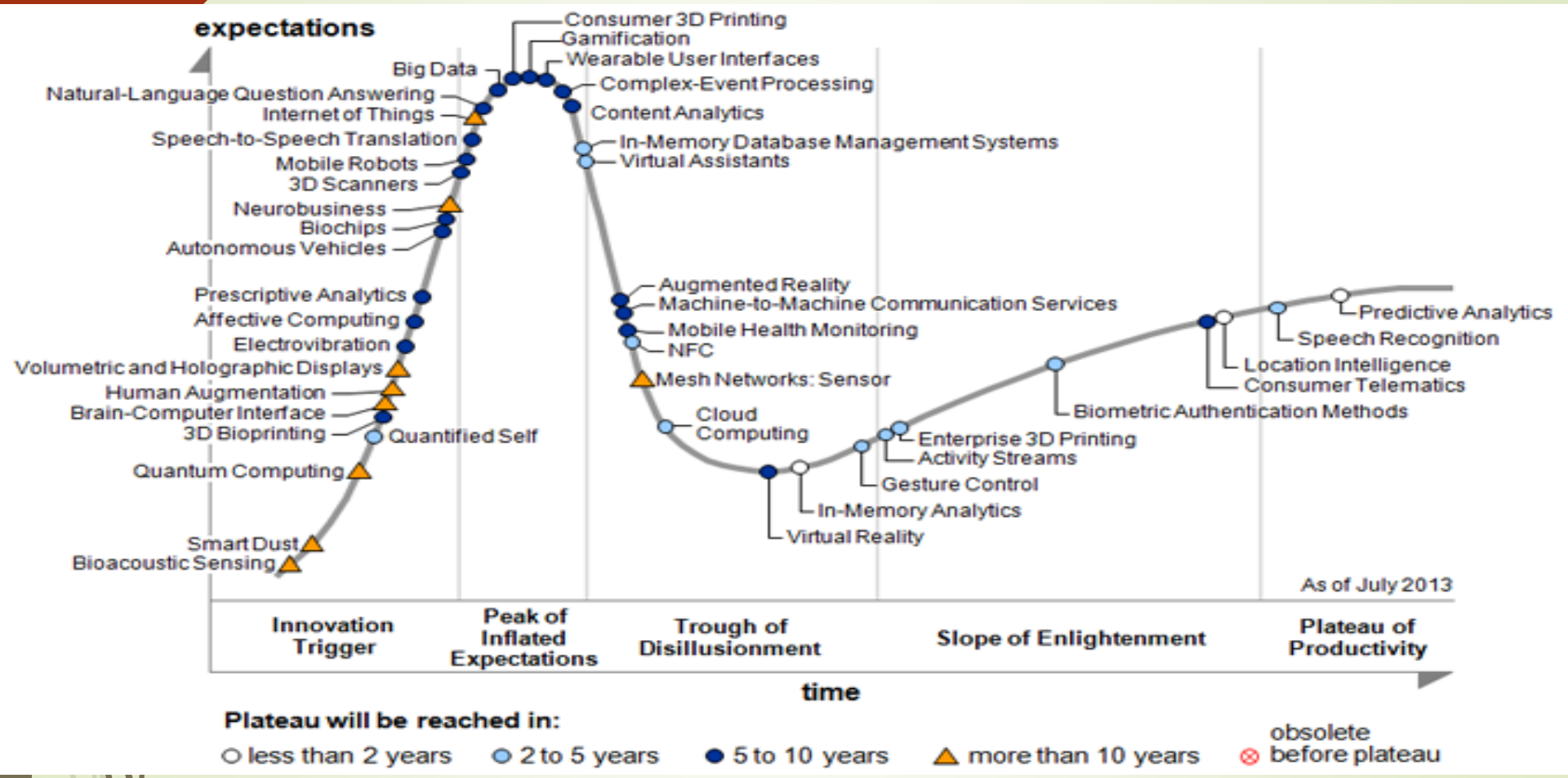
○ less than 2 years ○ 2 to 5 years ● 5 to 10 years △ more than 10 years ⊗ obsolete before plateau

Hadoop will be in most advanced analytics products by 2015, Gartner says

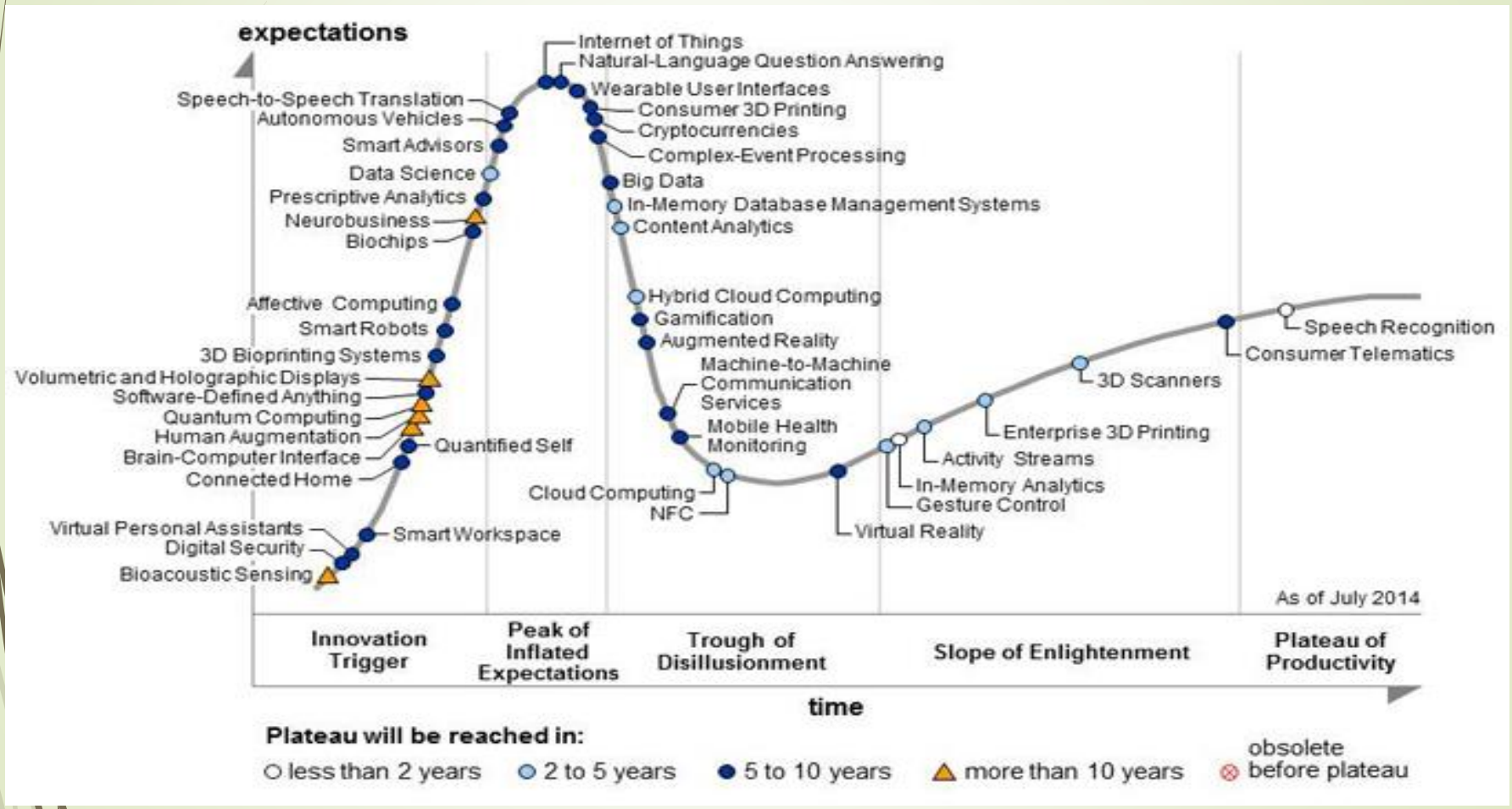
Технология Big Data (2012)



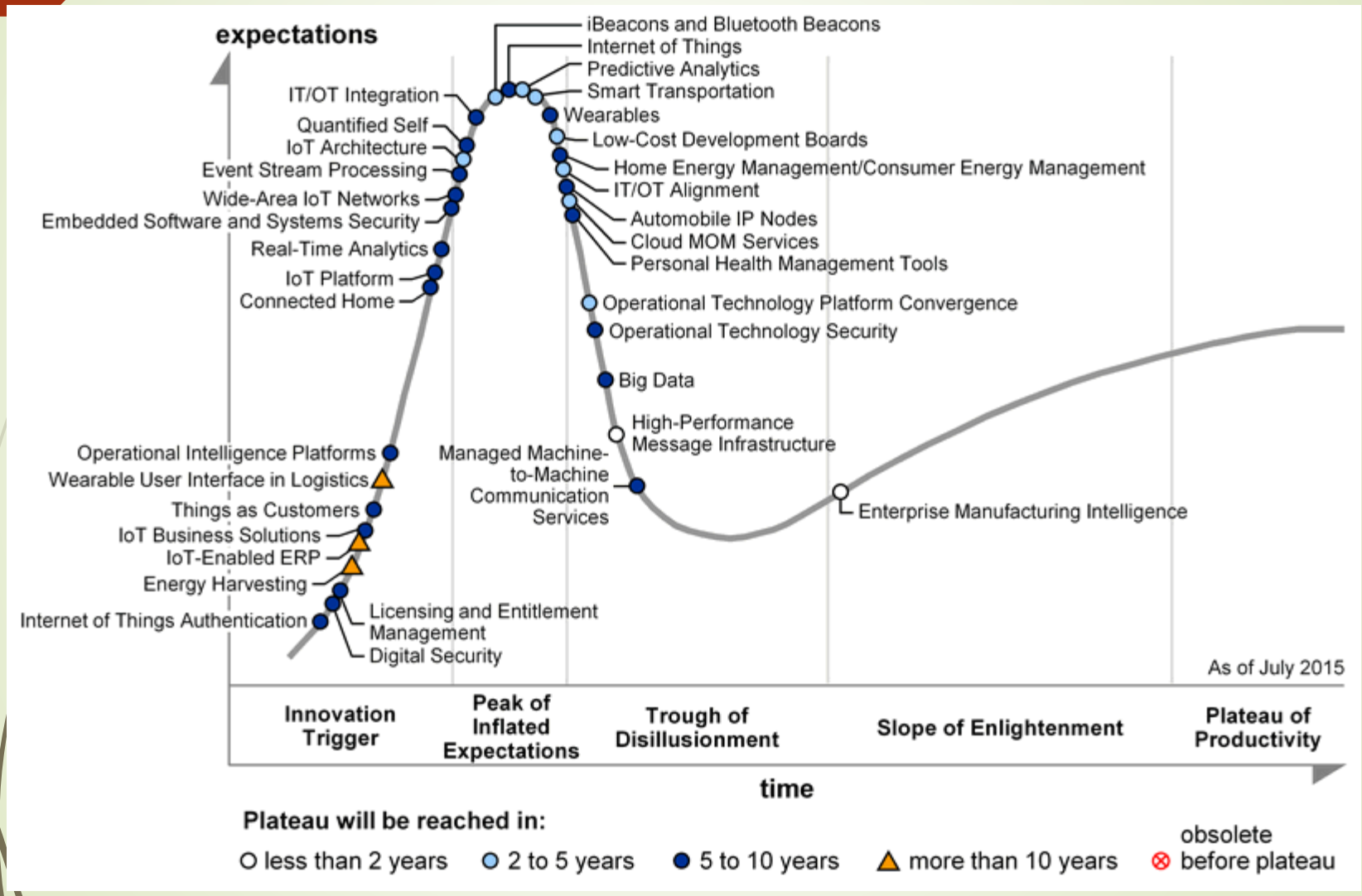
Технология Big Data (2013)



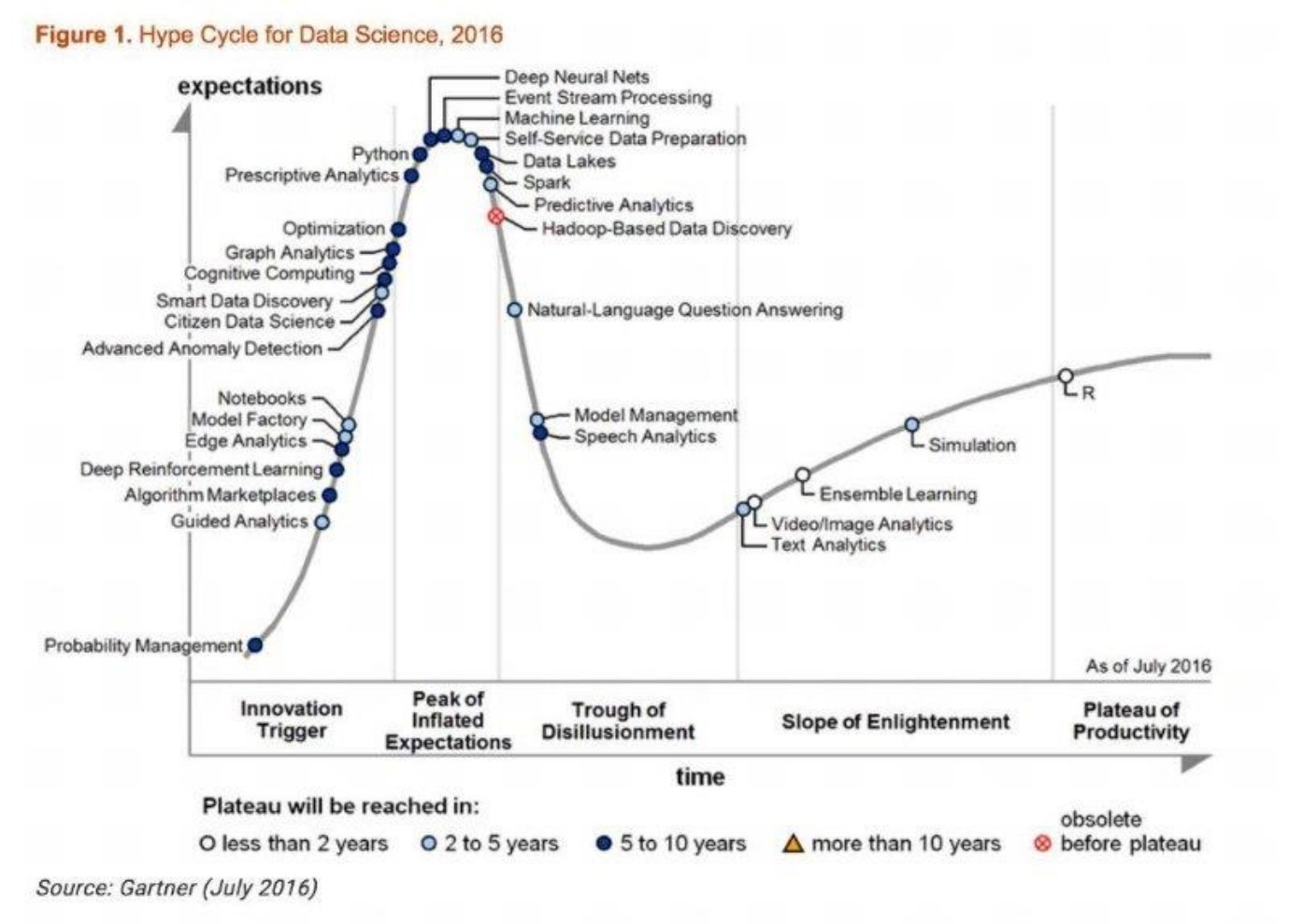
Технология Big Data (2014)



Технология Big Data (2015)



Технология Big Data (2016)



Big data is no where to be seen the hype cycle 2016 where as last year it is shown to enter trough of disillusionment. This may mean that the most talked about big data related technologies are now into practice and no more a hype.

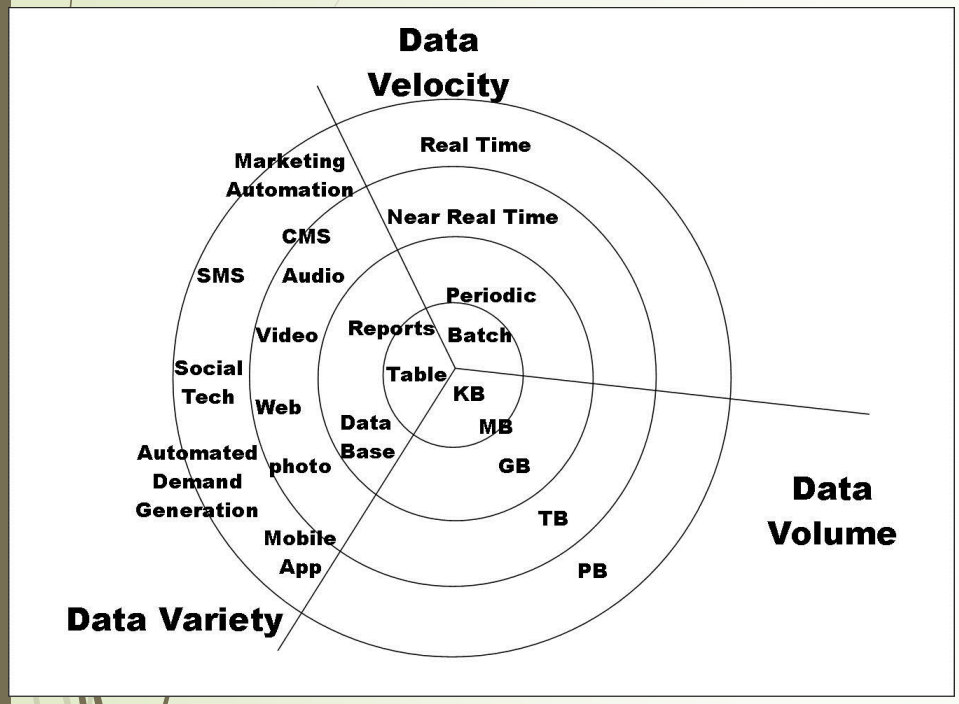
Big Data: Терминология

- Введение термина **«большие данные»** (Big Data) относят к Клиффорду Линчу, редактору журнала Nature, подготовившему к 3 сентября 2008 года специальный выпуск с темой «*Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объёмами данных?*»

- в котором были собраны материалы о феномене **взрывного роста объёмов** и **многообразия обрабатываемых данных** и технологических перспективах в парадигме вероятного скачка «от количества к качеству»;

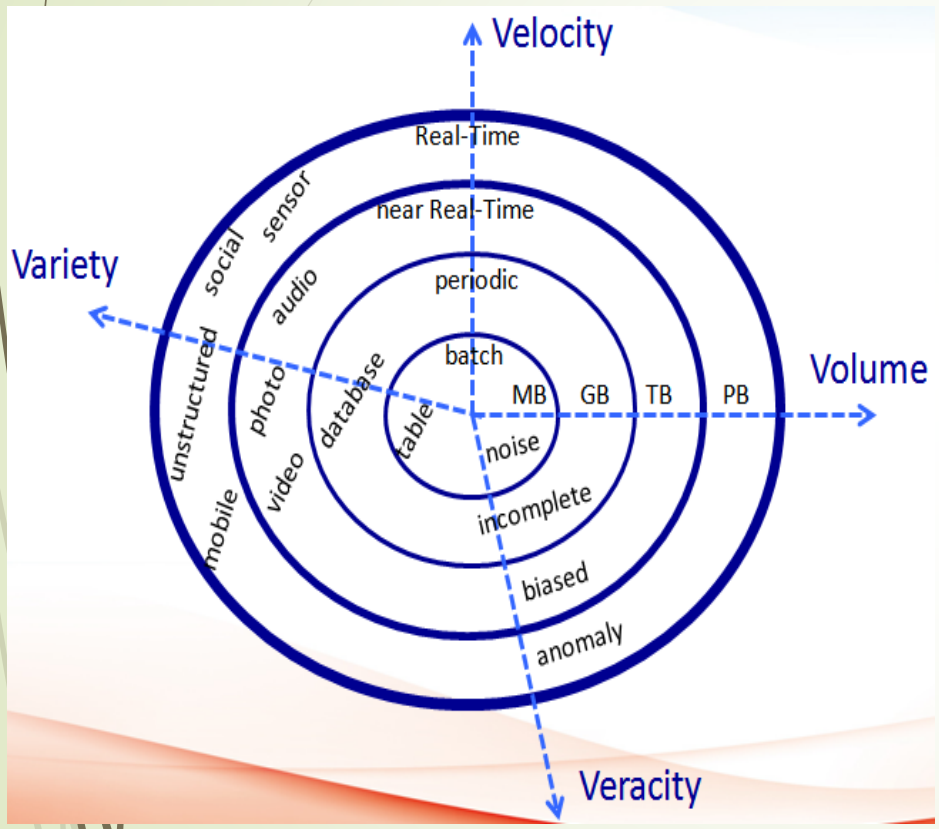
- термин был предложен по аналогии с расхожими в деловой англоязычной среде метафорами «большая нефть», «большая руда»

3v - Big data



- **Объём:** относится к наборам данных, размер которых выходит за пределы возможностей программных средств типичной базы данных собирать, хранить, обрабатывать и анализировать данные
- **Разнообразие:** способность обработки множества типов, источников и форматов данных от сенсоров, умных устройств, социальных сетей. Интегрировать все большее число источников, содержащих различные данные (структурированные наряду с сырыми, слабоструктурированными, неструктурированными данными, извлекаемыми из web- страниц, web log файлов, e-mails, документов и др.)
- **Скорость:** реакция на текущую информацию за время, ограниченное приложением; потоковая обработка (например, GPS данных в реальном времени). Приложения: электронная торговля, размещение на Web страницах в реальном времени, определение намерений пользователей в реальном времени, мобильные социальные сети

4V,5V - BiGdata



Veracity - Достоверность

Имея дело с большим объемом, скоростью поступления/изменения и разнообразием данных, надо быть готовым к тому, что не все данные будет стопроцентно достоверны. Вопрос состоит в том, какая «чистота» данных достаточна для верного анализа? Зачастую данные и не должны быть «совершенными», однако они должны максимально приближаться к требуемому уровню достоверности. В зависимости от сферы применения, верификация данных может иметь важное значение либо быть простой опцией.

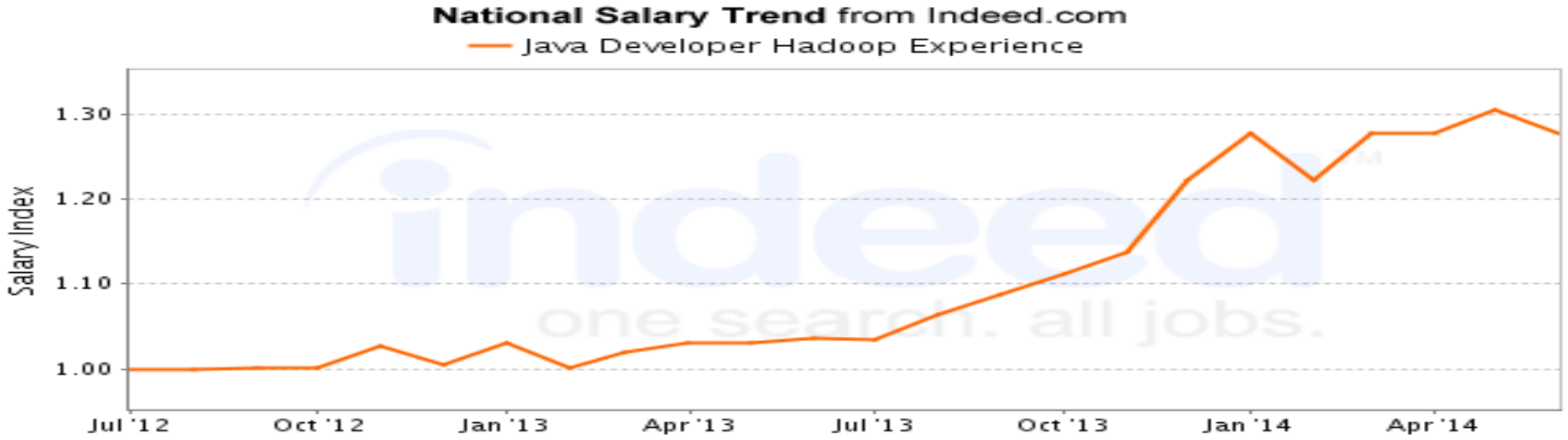
Value - Стоимость

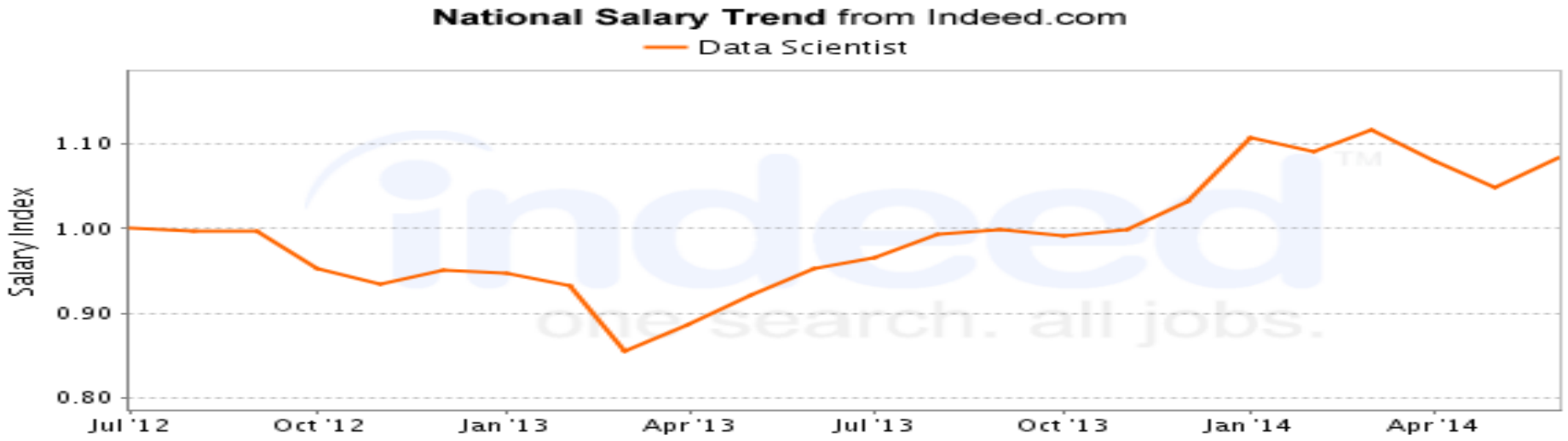
Это самый важный аспект больших данных. Реализация ИТ-инфраструктуры систем хранения и обработки больших данных стоит достаточно дорого, и бизнес будет требовать возврата инвестиций. В конце концов, если компания не может извлечь выгоду из своих данных, нет никакого смысла в наращивании потенциала для хранения и управления ими.

Дефицит профессионалов для анализа данных

- Данные в мире увеличатся к 2015 г. до более чем 7.9 зетта байт
- Предприятия эффективно используют менее 5% доступных данных, главным образом, из-за нехватки специалистов, способных профессионально иметь дело с массивными объемами данных (Форрестер)
- Многие правительства понимают необходимость Инициативы Big Data, поскольку можно экономить 100 млрд. долларов в год на административных расходах (Мак Кинзи)
- Рабочих мест для образованных, тренированных и опытных в работе с данными профессионалов будет становиться все больше
- Нехватка 140 000 - 190 000 рабочих мест может удвоиться, или утроиться в течение следующих 20 лет
- Университеты и высоко технологичные компании работают совместно, чтобы восполнить этот разрыв

Big Data работа (hadoop, data scientis)





Типичный набор компетенций, требуемый для занятия вакансии

- Создание моделей данных для анализа конкретных бизнес-ориентированных вопросов
- Организация, анализ, поиск, классификация, отчеты по Big data
- Проектирования и использование средств Machine Learning и других методов data mining
- Разработка продвинутых методов визуализации данных
- Конструирование прогностических и статистических моделей
- Создание масштабируемых процессов и метрик для принятия деловых решений
- Презентации на любом уровне организации, включая коллег, управляющий персонал, клиентов организации
- Контроль полного цикла обработки данных ETL
- Применение технологий NoSQL подобных Pig, Hive, Native M/R для реализации произвольного анализа данных

Типичная требуемая квалификация

- Магистр (Math, Statistics, CS or physics preferred). Предпочтительнее Ph.D
- Необходим опыт программирования в R или в других статистических пакетах. Знание Python и Java – это плюс
- Опыт со средствами визуализации и методами (i.e. Tableau, QlikView D3, etc.)
- 5+ летний опыт работы с широким набором статистических задач и методов
- 4+ летний опыт конструирования алгоритмов и аналитических приложений
- Прекрасная школа коммуникации, включая способность перевода и презентаций
- Инновационное, творческое, проактивное мышление
- Опыт работы с реляционными базами данных и системами поддержки Big Data, такими как Hadoop, HBase, AWS

Объявление Белым Домом деталей новой стратегии Big Data

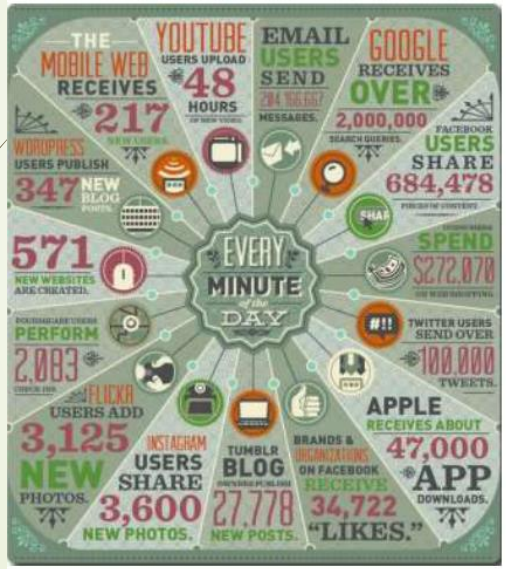
- **Пресс Релиз (29 марта, 2012):** “Подобно прежним Федеральным вложениям в R&D информационных технологий, которые привели к огромным достижениям в области суперкомпьютеров и создании Интернета, объявляемая сегодня инициатива нацелена на трансформацию нашей способности использовать Big Data для научных открытий, биомедицинских исследований и исследований окружающей среды, образования, и национальной безопасности

- **Гранты:** Около **$73** миллионов будет специально вложено в исследовательские гранты, из которых **National Science Foundation** выделит **$13** миллионов на три проекта, а Министерство Обороны выделит до **$60** миллионов.

Примеры областей применения Big Data

- **Социальные и гуманитарные науки** (пример проекта: Антипреступная и антикоррупционная обсерватория)
- **Науки об окружающей среде** (пример проекта: Сейсмологические проблемы для понимания динамики Земли в рамках Пан-Европейской инфраструктуры управления океаническими данными)
- **Биомедицинские науки** (пример проекта: Распределенное хранение и анализ больших многомерных данных об энцефалограммах на основе Hadoop и Hbase)
- **Материаловедение** (пример проекта: Qnano – Пан-Европейская инфраструктура для оценки качества при тестировании наноматериалов)
- **Физические науки** (пример проекта: LHC – комплекс Большого Адронного Коллайдера)
- **E-Инфраструктуры** (пример проекта: Обработка траекторных сенсорных данных при помощи MapReduce)
- **Энергетика** (пример проекта: RIEEB – Исследовательские инфраструктуры для эффективного использования энергии в зданиях)

Внедрение



Когда появилась big data?
технологии Apache Hadoop - +/-7 лет.

Hadoop – open source экосистема вокруг больших данных.
Пока эти технологии считаются незрелыми, находящимися на начальной стадии проникновения в бизнес-среду.
Hadoop позволяет фильтровать, сортировать и хранить данные, а также обрабатывать огромные их объемы (неструктурированные)

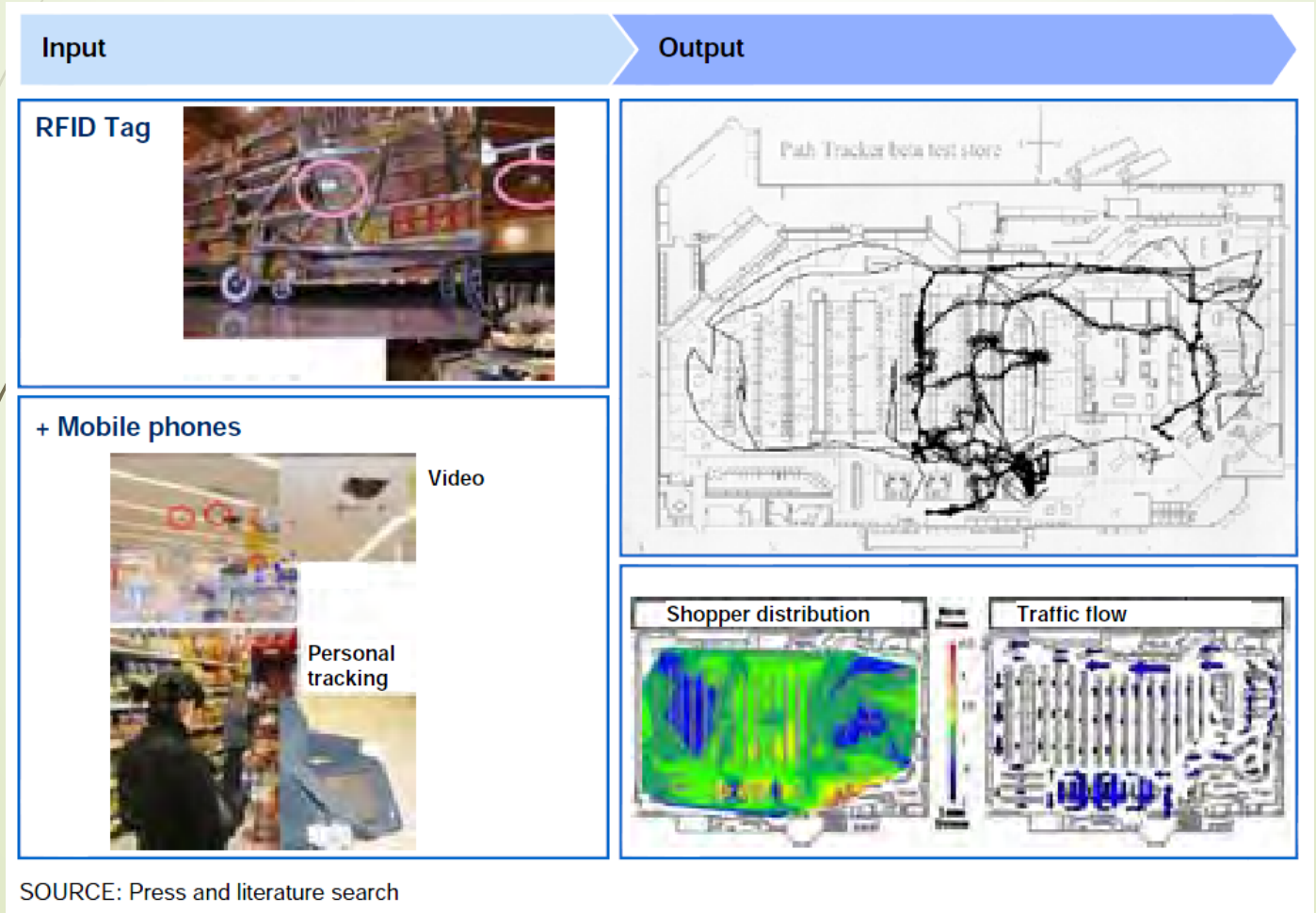
Hadoop внедрена в Bank of America, Merrill Lynch
(Основные пользователи технологий Big Data - инвестиционные банки)

Примеры

- Расшифровка генома человека заняла 10 лет. Теперь это можно сделать за 1 неделю.
- Проанализировав результаты запросов Google, ученые обнаружили, что в развитых странах люди чаще интересуются будущим, чем прошлым.
- Правительство Б. Обамы развернуло 84 программы обработки больших данных. По оценке политиков результаты этих программ помогли Б. Обаме победить на выборах в 2012 г.
- Размер базы данных NASA результатов наблюдения за климатом занимает 32 пегабайт и обрабатывается на суперкомпьютерном кластере.
- В 2005 г. базы данных Amazon.com составляли около 8 Тб
- Walmart осуществляет более 1 млн. транзакций каждый час. База данных содержит 2.5 пегабайт, что по объему в 167 раз больше объема библиотеки Конгресса США.
- База данных Facebook содержит 50 млрд. фотографий.

Big Data Analysis Example: Product arrangement

- How does location tracking work?
  - Recognize the dead zone



SOURCE: Press and literature search

Facebook

IBM Watson

- В феврале 2011 года для проверки возможностей Уотсона он принял участие в телешоу Jeopardy! (российский аналог — Своя игра)[3]. Его соперниками были Брэд Раттер — обладатель самого большого выигрыша в программе, и Кен Дженнингс — рекордсмен по длительности беспроигрышной серии. Уотсон одержал победу, получив 1 миллион долларов, в то время, как Дженнингс и Раттер получили, соответственно, по 300 и 200 тысяч.
- Уотсон состоит из 90 серверов Power7 750, каждый из которых содержит по 4 восьмиядерных процессора POWER7. Суммарная оперативная память Уотсона более 15 терабайт.
- Система имела доступ к 200 миллионам страниц структурированной и неструктурированной информации объемом в 4 терабайта,[4] включая полный текст Википедии. Во время игры Уотсон не имел доступа к интернету.

IBM Watson в медицине

- Шесть экземпляров IBM Watson уже «трудоустроены» в больницы США в качестве врачей-диагностов (на начало 2013)
- За два года учёбы Watson изучил 605 тыс. медицинских документов, в общей сложности 2 миллиона страниц текста
- Перед началом врачебной практики компьютер проанализировал 25 тыс. историй болезни и проработал 14,7 тыс. для тонкой настройки алгоритмов
- Watson способен не только ставить диагнозы, но и определять наиболее оптимальный курс лечения
- Точность назначения оптимального лечения после диагностирования рака лёгких в больницах США составляет 50%. То есть в половине случаев врачи рекомендуют не самый идеальный курс лечения и препаратов.
- У компьютера IBM Watson точность назначения оптимального лечения составляет 90%.

IBM Watson в колл-центрах

- В США на работу колл-центров служб поддержки пользователей ежегодно тратится 112 миллиардов долларов
- половина из 270 миллиардов обращений остаётся нерешённой
- По данным IBM, почти две трети из нерешённых вопросов клиентов могли бы быть решены, если бы сотрудники колл-центров могли искать информацию быстрее
- В ближайшие месяцы тестирование Watson в службах поддержки начнут первые клиенты IBM
  - австралийский банк ANZ, Royal Bank of Canada, Nielsen, занимающаяся исследованием потребительского поведения, израильский мобильный оператор Celcom, консалтинговая и аналитическая компания IHS
- IBM уже тестирует Watson в собственных колл-центрах, и первые результаты впечатляют — компьютер сообщает клиенту нужную информацию на 40% быстрее

Технологии Big Data

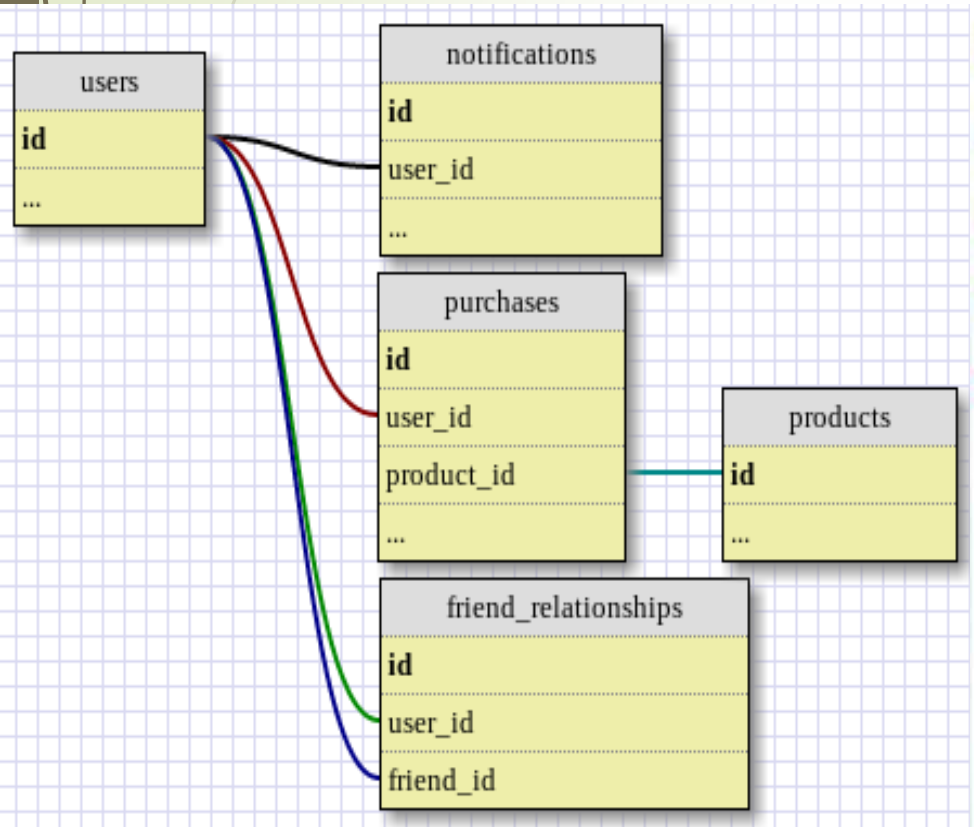
- Shared Nothing Architecture
- NoSQL
- MapReduce, Hadoop
- R (язык программирования)

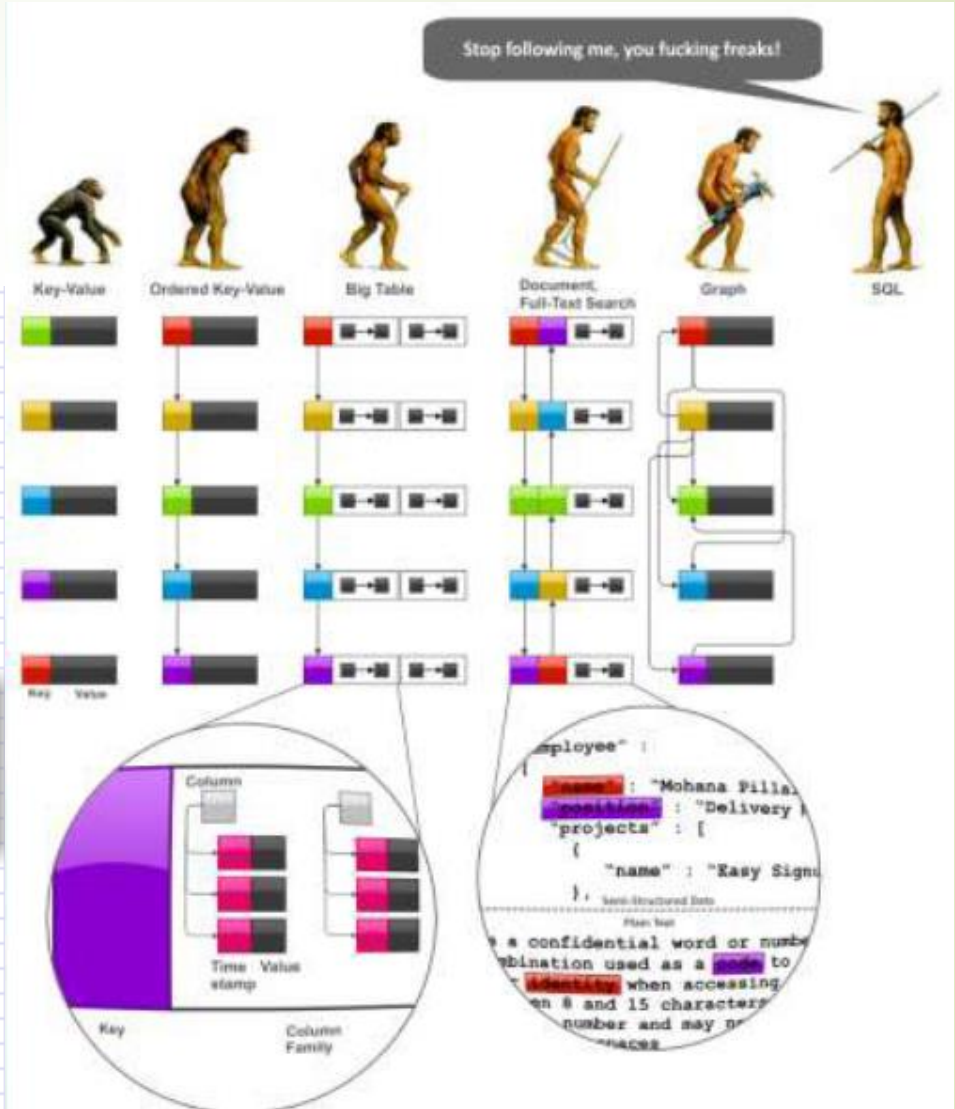
SNA

- Shared nothing architecture (SNA) – архитектура независимых распределенных вычислений, в которой отдельные узлы имеют собственную память, дисковые массивы и устройства ввода/вывода. Каждый узел в такой архитектуре самодостаточен и ничем не делится с другими узлами сети.
- Такая архитектура хорошо масштабируется и становится все более популярной.
- Каждый узел в SNA выполняет собственную задачу, взаимодействуя с другими узлами по специальному протоколу

NoSQL

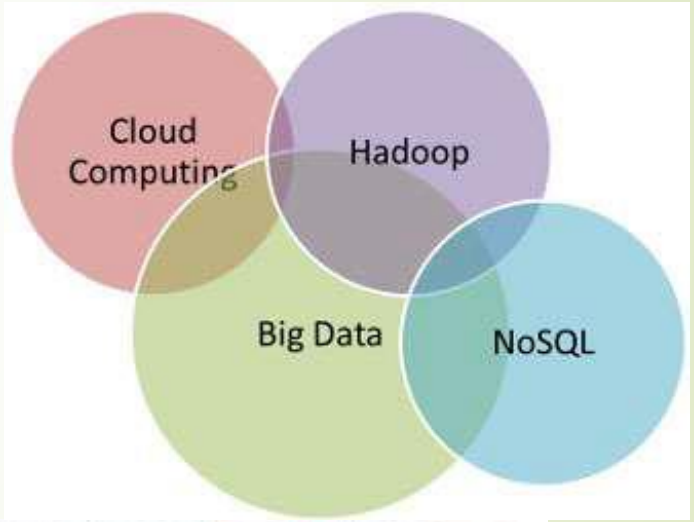
- Новый тип баз данных: нереляционные, распределенные, с открытым кодом и горизонтально масштабируемые

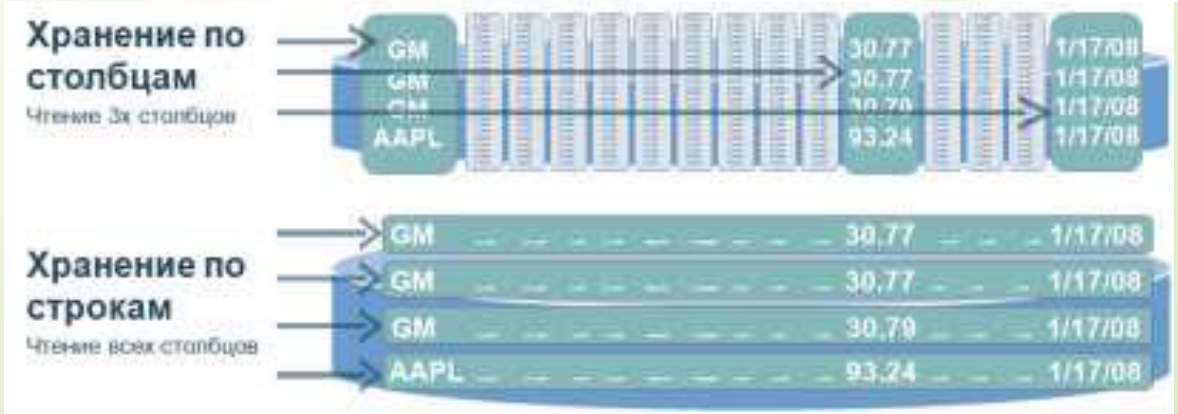




Новые виды баз данных

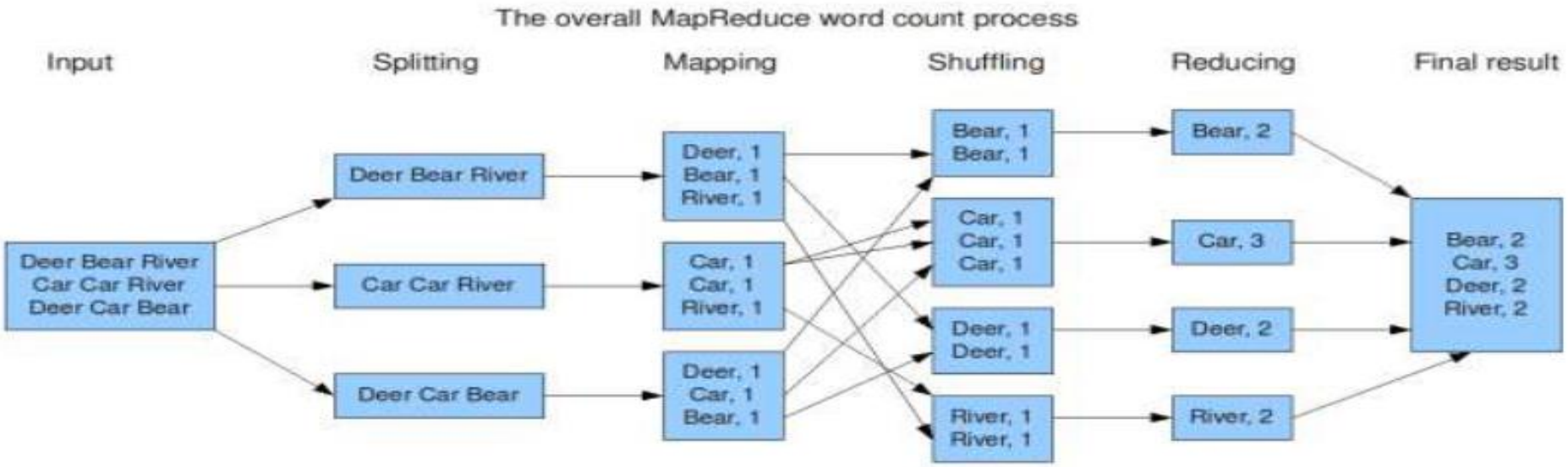
- Базы данных с встроенной аналитикой (DB Warehouse, Netezza, Teradata, …)
- In-memory базы данных (SAP HANA)
- Колоночные хранилища данных (Hbase)
- Графовые базы
- Документные базы (MongoDB)
- Комбинированные базы (DB2)





Map-Reduce

- Модель распределённых вычислений Google для параллельных вычислений очень больших, несколько петабайт, данных в компьютерных кластерах.
- Работа MapReduce состоит из двух шагов: Map и Reduce.
  - На Map-шаге происходит предварительная обработка входных данных. Для этого один из компьютеров (называемый главным узлом — master node) получает входные данные задачи, разделяет их на части и передает другим компьютерам (рабочим узлам — worker node) для предварительной обработки.
  - На Reduce-шаге происходит сбор предварительно обработанных данных. Главный узел получает ответы от рабочих узлов и на их основе формирует результат — решение задачи.

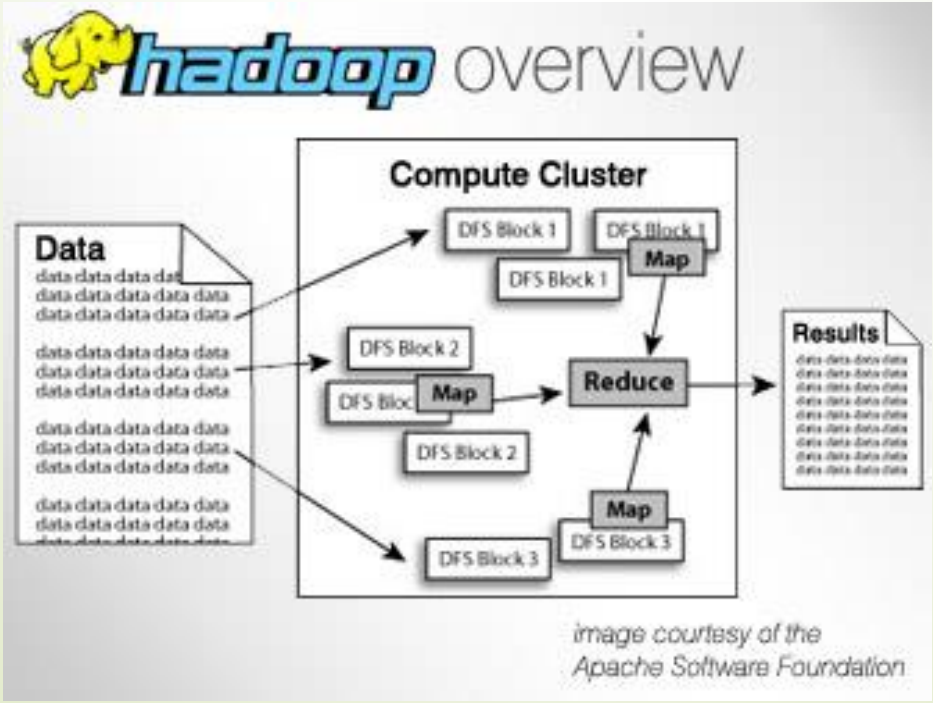


Преимущества Map-Reduce

- Средняя производительность HDD ~100МБ/с
- Прочесть 1 ТБ ~ 2.5 часа
- Прочесть 1 ТБ параллельно со 100 дисков ~ 2 минуты
- Произвольный доступ к диску медленный
- Последовательный доступ быстрый!

Hadoop

- Система хранения и обработки данных с открытым кодом
- Масштабируемая и распределенная
- Основана на работе Google
  - GFS + MapReduce + BigTable



Новый виток в развитии методов

- Искусственные нейронные сети
- Методы предиктивной аналитики
- Статистические методы
- Математическая лингвистика
- Краудсорсинг
- Сентимент-анализ
- Распределенные файловые системы

Девять Big Data платформ крупных фирм

1) Облачные варианты: **Amazon** предлагает Elastic Map Reduce (масштабируемый Hadoop), инструменты для аналитики, DynamoDB (NoSQL СУБД)

2) Облачный подход Google к Big Data: BigQuery — аналитическая база данных с прогнозным API, еще незрелое предложение, нуждающееся в средствах поддержки machine learning. Its cloud application hosting service, Google AppEngine также предлагает средства MapReduce

3) HP приобрел поколонную, ориентированную на Big Data аналитику **СУБД Vertica** и предлагает коннектор для Hadoop

4) Microsoft разрабатывает собственное основанное на **Hadoop** решение в **Windows Azure как облачный вариант**. Планируется **open source Hadoop on Windows**

5) Oracle объявил **Big Data Appliance в январе** 2012. В нем используется Cloudera-вариант Hadoop'а, а также Cloudera Manager, R + **Exadata, Exalytics, Analytics, BI, Clouds**

Девять Big Data платформ крупных фирм (2)

6) Подход SAP к Big Data основан на интеграции с другими инструментами. Для транзакционных данных предлагается **SAP HANA**. В области Big Data, подобно Hadoop, SAP выходит на рынок на основе **Sybase IQ, в котором добавлены существенные возможности Map-Reduce**

7) Teradata Aster предлагает программную платформу и аппаратурное решение, поддерживающие аналитическую обработку как в **SQL так и в MapReduce**

8) SAS работает над решением, соединяющим **Hadoop Distributed File System (HDFS)** с аналитикой в оперативной памяти

9) Big Data платформа IBM InfoSphere включает ряд компонентов: DB2 , Cloud , Hadoop (InfoSphere BigInsights), Hive, Pig, Jaql, Hbase, Netezza, средства анализа данных, и др.

Доход компаний поставщиков Big Data

Вендор	Доходы от больших данных ($ млн)	Общий доход ($ млн)	Доля доходов от больших данных в общей выручке
IBM	$953	$106,000	1%
Intel	$765	$54,000	1%
HP	$513	$126,000	0%
Fujitsu	$285	$50,700	1%
Accenture	$273	$21,900	0%
CSC	$160	$16,200	1%
Dell	$154	$61,000	0%
Seagate	$149	$11,600	1%
EMC	$138	$19,000	1%
Teradata	$120	$2,200	5%
Amazon Web Services	$116	$650	18%
SAS Institute	$115	$2,700	24%
Capgemini	$111	$12,100	1%
Hitachi	$110	$100,000	0%
SAP	$85	$17,000	0%

Вендор	Доходы от больших данных ($ млн)	Общий доход ($ млн)	Доля в общей выручке
Think Big Analytics	$8	$8	100%
MapR	$7	$7	100%
Digital Reasoning	$6	$12	50%
ParAccel	$5	$11	45%
Couchbase	$5	$6	84%
DataStax	$4.5	$4.5	100%
10gen	$4.5	$4.5	100%
Datameer	$4	$4	100%
Hortonworks	$3	$3	100%
RainStor	$2.5	$2.5	100%
Attivio	$2.5	$19	13%
QlikTech	$2	$300	1%
HPCC Systems	$2	$2	100%
Karmasphere	$2	$2	100%

Традиционные архитектуры

- Архитектура СУБД:
  - Дисковое хранение
  - Транзакции
  - Индексация
  - Журнализация
  - Многопоточность
  - Блокировки
- Производительность самых мощных серверов растет медленнее чем объемы данных

Традиционные архитектуры (недостатки)

- Оптимизация и индексы влекут накладные расходы на дополнительные операции, а следовательно сильное падение скорость при нагрузках
- Кеш на чтение влечет за собой усложнение клиентского ПО
- Очередность операций влечет проблему хранения и ограниченности очереди
- Горизонтальное масштабирование влечет денормализацию и ограничения на join
- Денормализация схемы влечет избыточность данных

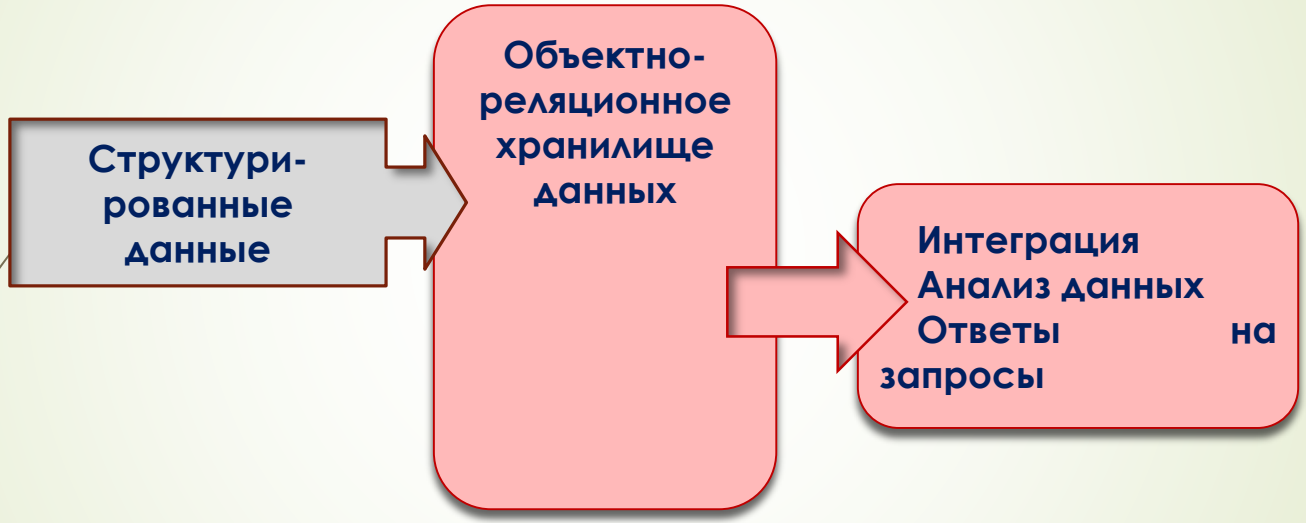
Традиционные архитектуры (пути решения)

- Отказ от строгой консистентности
- Уход от нормализации и внедрение избыточности
- Потеря выразительности SQL и моделирование части функций программно
- Усложнение клиентского ПО
- Сложность поддержания работоспособности и отказоустойчивости

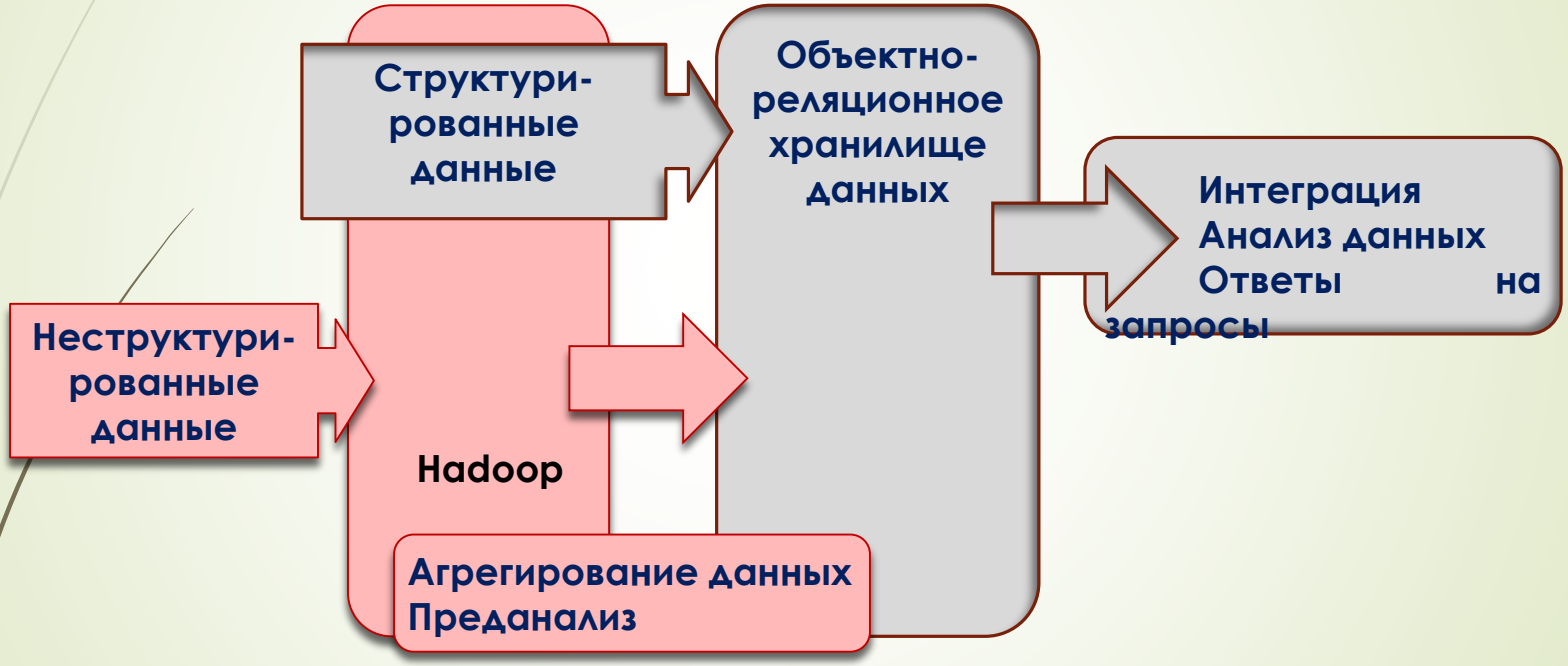
NoSQL

- Неограниченное горизонтальное масштабирование
- Произвольная схема вплоть до отсутствия таковой
- Согласованность данных невзирая на производительность
- Не большие трудозатраты на развертывание больших кластеров
- Императивные программы наряду а иногда и вместо декларативных программ

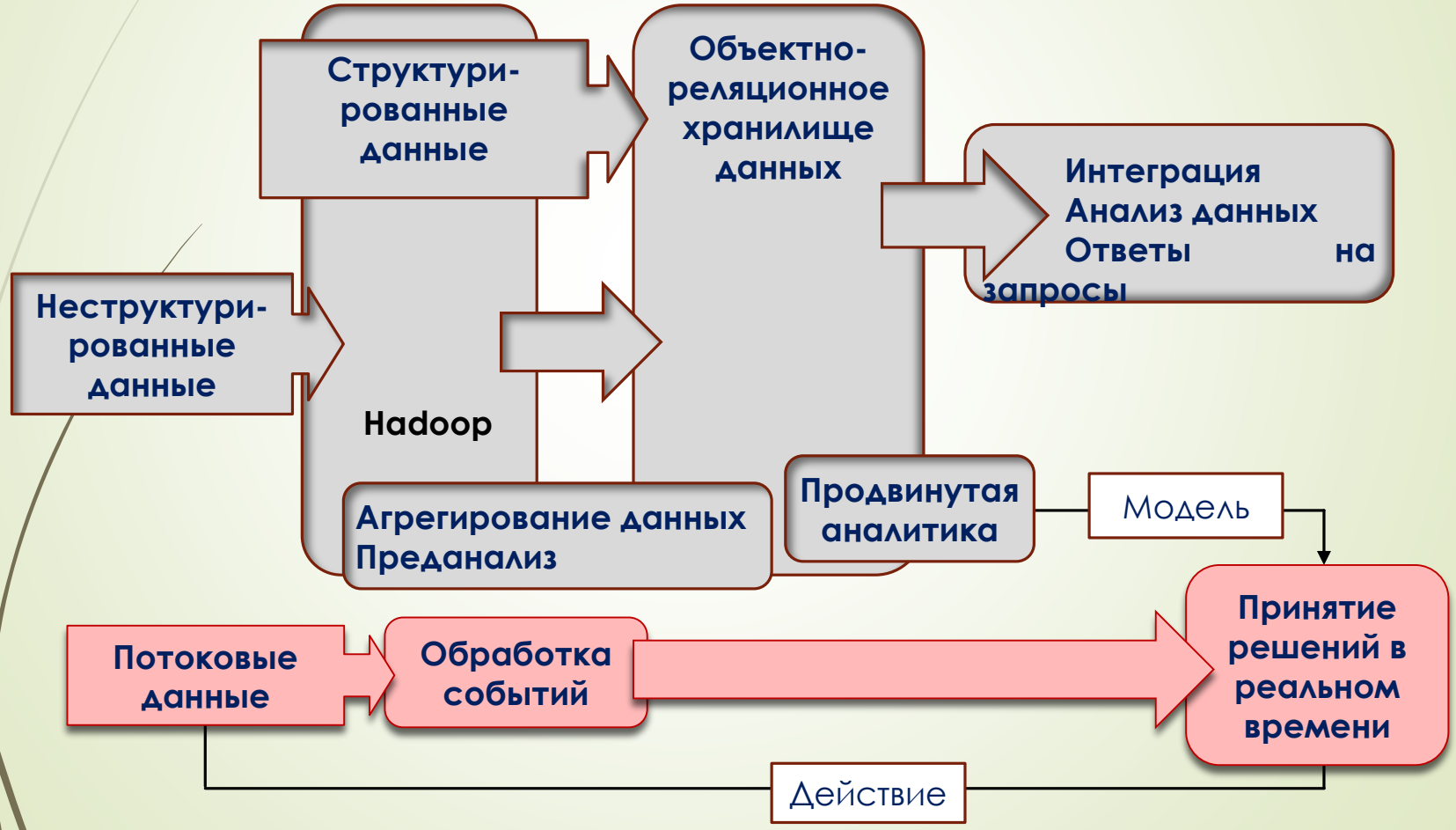
Компоненты традиционной архитектуры



Развитие архитектуры для поддержки Volume и Variety Big Data



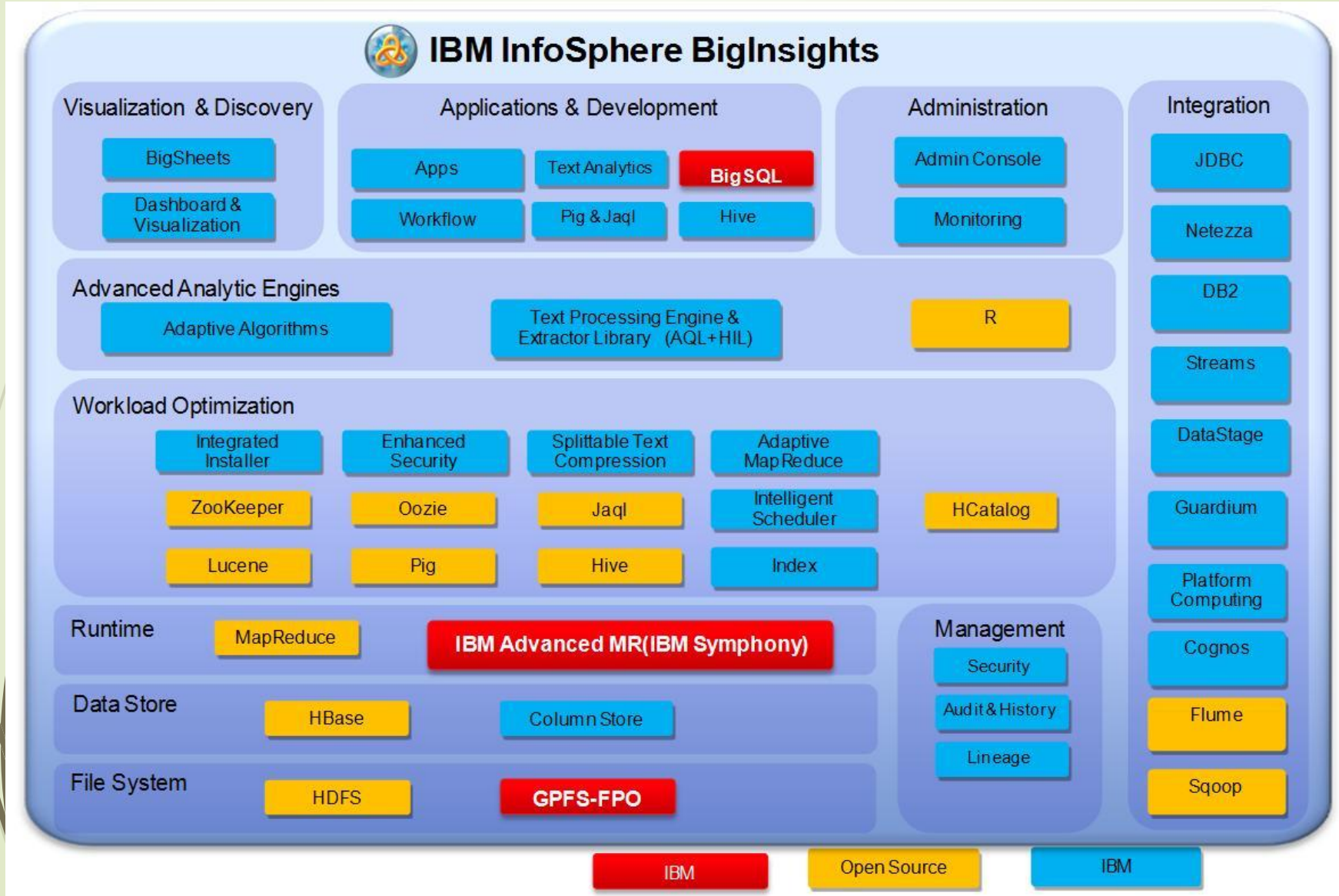
Развитие архитектуры для поддержки Velocity Big Data



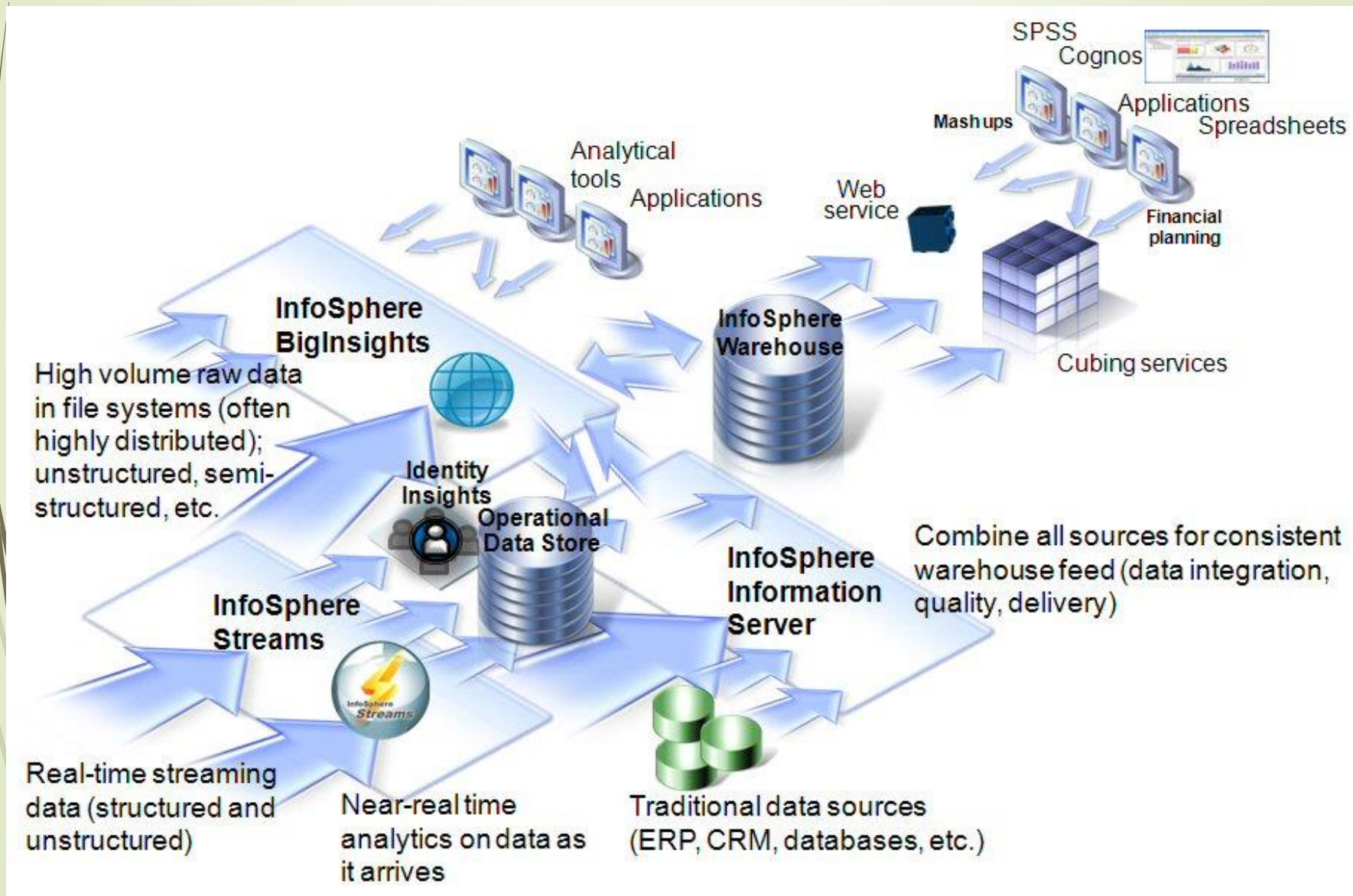
Средства обнаружения новой информации

Платформа IBM Big Data

InfoSphere BigInsights – A Full Hadoop Stack (IBM)



Аналитика в платформе IBM Big Data



Map Reduce vs Parallel DBMS

	Parallel DBMS	MapReduce
Schema Support	✓	Not out of the box
Indexing	✓	Not out of the box
Programming Model	Declarative (SQL)	Imperative (C/C++, Java, …) **Extensions through Pig and Hive**
Optimizations (Compression, Query Optimization)	✓	Not out of the box
Flexibility	Not out of the box	✓
Fault Tolerance	Coarse grained techniques	✓

План Курса

1. (Лекция) Введение в курс. Основные понятия курса. Введение в проблематику управления большими данными. Сравнение с традиционными подходами моделирования данных и манипулирования данными в СУБД
2. (Лекция) Введение в Hadoop, архитектура Hadoop, парадигма map-reduce, Архитектура стека компонентов Hadoop, Администрирование Hadoop, запуск приложений (Jobs), мониторинг приложений, Импорт/экспорт больших данных в/из Hadoop
3. (Семинар) Работа с HDFS, запуск простейших заданий (подсчет слов), администрирование Hadoop кластера, Импорт/экспорт больших данных в/из Hadoop

План Курса (2)

4. (Лекция) Алгоритмы известных операций в map-reduce (матрично-векторное умножение, реляционные операции и др.)
5. (Лекция) Разработка собственных приложений Map-Reduce, Потоки работ Map-Reduce
6. (Лекция) Теория сложности алгоритмов для map-reduce
7. (Семинар, Домашнее задание) Разработка на Java собственного приложения map-reduce на примере задачи подсчета слов, реализация одного из алгоритмов известных операций в map-reduce

План Курса (3)

8. (Лекция) Языки программирования высокого уровня над Hadoop: Pig, Hive, JAQL, BigSQL
9. (Семинар) Реализация практического задания на Pig, Hive, JAQL, BigSQL
10. (Лекция) Базы данных нетрадиционных моделей данных Not Only SQL (NoSQL) и их связь с Hadoop
11. (Семинар) Реализация практической задачи с помощью HBase

План Курса (4)

12. (Лекция) Интеграция Hadoop с традиционными хранилищами данных (базами данных и warehouse)
13. (Семинар) Освоение импорта, экспорта данных из/в Hadoop и DB2 Warehouse (Netezza)
14. (Лекция) Взаимодействие методов анализа данных с Hadoop
15. (Семинар) Реализация задач с помощью RHadoop